Patterns of polymorphism, selection and linkage disequilibrium in the subgenomes of the allopolyploid *Arabidopsis kamchatica*

Supplementary Material

Reference genome assembly of *A. lyrata* subsp. *petraea*

We assembled the genomes of *A. halleri* subsp. *gemmifera* (W302) collected from Tada mine in Japan and *A. lyrata* subsp. *petraea* (lyrpet4) collected from Siberia representing each of the closest known diploid parents of *A. kamchatica* [1,2]. Both *A. halleri* and *A. lyrata* are predominantly self-incompatible (SI). To reduce heterozygosity, we selfed *A. halleri* five times using bud pollination [3]. The Siberian *A. lyrata* genotype (lyrpet4) lost SI in its natural habitat, so we were able to perform two rounds of regular self-fertilization. Previously, we reported medium quality assemblies (v1.0) for both of these genotypes [4] as well as an improved version of *A. halleri* [5]. Here, we provide an improved version of the *A. lyrata* lyrpet4 assembly that was generated using the pipeline described in Briskine et al. [5] for *A. halleri* W302 and we refer to the new assemblies as version 2.2 (v2.2).

We created long-insert mate-pair libraries to complement the short-insert libraries published in Akama et al. (2014). We used selfed *A. lyrata* lyrpet4 leaf tissue to construct the mate-pair libraries with Illumina Nextera Mate-Pair Library Prep kit modified for large insert sizes. After tagmentation with Mate Pair Tagment Enzyme, the DNA was separated by pulse field electrophoresis into variable ranges of 22-38 kb, 15-22 kb, 11-15 kb, 7-11 kb, 5.0-7 kb, and 3.0-5.0 kb. For each range, 270 – 600 ng of DNA was recovered using a Zymoclean Large Fragment DNA Recovery Kit. After circularization, exonuclease treatment, fragmentation with Covaris S1, A-tailing, and adapter ligation, 14 cycles of PCR were carried out for 22-38 kb, 15-22 kb, and 11-15 kb fraction, and 10 cycles for the 7-11 kb, 5.0-7kb, and 3.0-5.0 kb fractions. After quantification of the libraries by qPCR using KAPA Library Quantification Kit for Illumina platforms, 4 additional cycles of PCR were performed for the 22-38 kb and 7-11 kb fractions. The libraries were purified with an AMpure XP kit, quantified with the KAPA kit again, and mixed based on the measurement. The libraries were sequenced on Illumina HiSeq 2500 at the Functional Genomics Center Zurich (http://www.fgcz.ch).

The *A. lyrata* genome was assembled from all available untrimmed read libraries with ALLPATHS-LG R50599 [6] using the default parameters in two steps. In the first step, we specified expected insert sizes. In the second step, we switched to the insert sizes reported by ALLPATHS-LG in the first step. The assembly job had a peak memory utilization of 191 Gb and completed in 84 hours on a Linux server using 30 cores.

Code for *A. lyrata* genome assembly
https://gitlab.com/rbrisk/AlyrAssembly

Genome annotation of *A. lyrata* subsp. *petraea*

Both parental genomes were annotated using the same pipeline based on the recommendations from the AUGUSTUS Development Team[7]. The details for *A. halleri* can be found Briskine et al. [5]. Here, we provide a brief description of the *A. lyrata* lyrpet4 annotation process (see pipeline flowchart in Briskine et al. [5]. First, we aligned un-stranded paired-end 100 bp reads from *A. lyrata* W1739_L2 (leaf) and W1739_R0 (root) libraries from Paape et al. [8] against the *A. lyrata* lyrpet4 assembly using STAR v2.4.0i [9]. We extracted intron hints from the alignments and combined them with *nonexonpart* hints obtained from the RepeatMasker v4.0.5 [10] output. The combined hints were supplied to AUGUSTUS v3.0.3 for the initial run. These obtained gene models were used to extract exon-exon junction sequences against which we aligned the original RNA-seq reads using bowtie2 v2.2.4 [11]. We merged exon-exon junction alignments with the alignments to the complete reference genome and used the combined data to produce intron hints for the final AUGUSTUS run. Human readable functional descriptions were added using the AHRD tool [12]. Reciprocal best BLAST hits were calculated individually between *A. halleri* W302 or *A. lyrata* lyrpet4 and *A. thaliana* TAIR10 by aligning all coding sequences using NCBI BLAST+ v2.2.29 and comparing the scores for hits longer than 200 bp. Similarly, we calculated reciprocal best BLAST hits between W302 or lyrpet4 and *A. lyrata* subsp. *lyrata* annotation v2.0 of strain MN47 v1.07 released by [13] for the Joint Genome Initiative (JGI) reference genome v1.07.

Improving diploid assemblies using synteny

Both *A. halleri* and *A. lyrata* diverged recently [2,14] and each has 8 chromosomes [15] allowing us to use the *A. lyrata* subsp. *lyrata* strain MN47 v1.07 reference assembly [16] to perform genome-wide synteny analysis. The complete genome, coding sequences, and gene annotation of *A. lyrata* JGI were downloaded from the Phytozome v9.0 website (http://phytozome.jgi.doe.gov). Coding sequences of *A. lyrata* JGI were aligned to our *A. lyrata* lyrpet4 assembly using BLAT v3.5 [17] with default parameters except maximum intron size. Because the longest intron in the *A. lyrata* lyrpet4 assembly was 44,703 bp, we set the maximum intron size to 50 kb. Hits were filtered, sorted, and merged into syntenic regions using custom Perl scripts (see the GitLab repository). We only considered the hits covering at least 85% of the query sequence and accepted the hit from a syntenic gene even when it did not have the highest score for the locus. If an *A. lyrata* lyrpet4 scaffold contained two neighboring loci that were syntenic to two *A. lyrata* JGI regions located on different chromosomes or more than 100 kb apart, the scaffold was split into two parts by removing the sequence of unknown nucleotides. Scaffolds were only split if the sequence of unknown

nucleotides N's at the cut site spanned at least 50 bp. After this correction, the scaffolds were sorted by length in descending order and named sequentially beginning with scaffold_1. Because *A. kamchatica* is a self-compatible species, we were able to remove most heterozygosity by self-fertilization and we treated the subgenomes as haploid (i.e. 8 homozygous chromosomes in each subgenome).

We were interested in comparing homeolog diversity in genes surrounding the *HEAVY METAL ATPASE 4* (*HMA4*) locus. By comparing synteny across multiple *Arabidopsis* species, we can increase the likelihood that homeologs separated by long distances or on multiple scaffolds are indeed syntenic in *A. kamchatica*. We used the published genomes of *A. lyrata* MN47 v1.07 and *A. thaliana* (TAIR, https://www.arabidopsis.org/) which are assemblies of entire chromosomes. We then compared both *A. halleri* W302 and *A. lyrata* v2.2 lyrpet4 regions surrounding the *HMA4* coding sequences. The heavy metal transporter *HMA4* contains three tandemly duplicated ATPase coding sequences in European and Asian *A. halleri* [5,18]. Following synteny analysis of the *HMA4* region, we are now able to examine genetic diversity over longer, contiguous, scaffold regions containing genes flanking *HMA4* coding genes to compare with the genomic background and between subgenomes. We centered the main genomic region containing the *HMA4* genes (3 tandem copies in *A. halleri* and *halleri*-origin homeologs and a single *lyrata*-origin copy in *A. kamchatica*) which we call "HMA4-M". The region spans 304 kb on H-origin scaffold_116 and 155 kb in L-origin scaffold_52. We then used the upstream (left-side) adjacent region ("HMA4-L", 125 kb for H-origin region and 183 kb in L-origin region), and the downstream adjacent (right-side) region ("HMA4-R", 105 kb in H-origin s0273 region and > 50 kb for L-origin region s0270). We also made alignments for the 118 genes in Fig. 2 in the main document with putative roles in metal tolerance, hyperaccumulation, metal ion transport, metal homeostasis were collected from the following resources:[18–24].

Reference assembly statistics

Our new *A. lyrata* assembly reduced the number of scaffolds from 281,536 from a previous version (v1.0, reported in [4] to 1675 in version 2.2. The genome sizes of our diploid genome assemblies are 196 Mb (of which 78.9 Mb is genes) for *A. halleri* and 175 Mb (of which 75.4 Mb is genes) for *A. lyrata* (Table 1, main text). Using flow cytometry, we estimated the genome size of *A. halleri* to be 250 Mb and for *A. lyrata* it is 225 Mb, indicating that our assembled genomes captured 78% and 77% of the total genomes of both species respectively. Using flow cytometry, we estimated a genome size of 460-480 Mb for *A. kamchatica* (with some variation between genotypes), indicating that the combined genome sizes of both diploids are very close to flow cytometry estimates for the allopolyploid.

The number of annotated genes in the *A. lyrata* v2.2 assembly (31,232) is similar to the number in our *A. halleri* (Tada Mine) v2.2 assembly (32,553), and to previously published *A. lyrata* subsp. *lyrata* (Hu et al. 2011) and *A. thaliana* gene annotations (Supplementary Table 1). Using reciprocal Blast hits (RBH) to determine orthology of the annotated gene models to *A. thaliana,* we found 21,433 *A. halleri* and 21,472 *A. lyrata* genes could be assigned to a TAIR10 gene ID. Based on these results, we identified 23,529 *halleri*-origin and *lyrata*-origin homeologs (Supplementary Table 2). Our *A. halleri* and *A. lyrata* v2.2 genome assemblies also show very similar numbers of blast hits to the JGI *A. lyrata* genome (Supplementary Table 3).

Population Structure

To estimate population structure and phylogenetic relationships in our *A. kamchatica* accessions, we used SNPs from coding regions for a thousand randomly sampled *halleri* (22,896 SNPs) and *lyrata* (23,637 SNPs) homeologs. When we combined the SNPs for both subgenomes into a single dataset, the highest support for STRUCTURE[25] assignment was $K = 2$ [26], consistent with previous analyses[1,27]. When subgenomes were analyzed separately, different clusterings were observed. The highest supported structure assignments were $K = 4$ for the *halleri*-subgenome and $K = 3$ for the *lyrata*-subgenome (Supplementary Fig. 2).

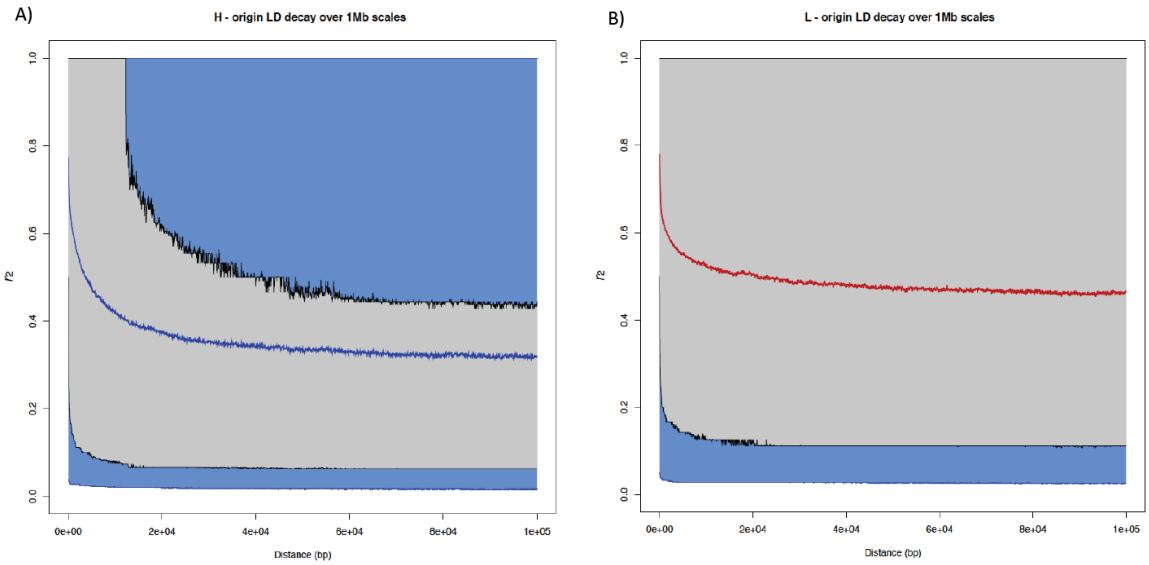
Phylogenetic relationships of the 25 accessions were consistent with population structure clustering described above. In each of the three phylogenies, three clades are fairly well resolved: one large clade from the southern species range (most of Japan), another main clade from the northern range containing samples from Far East Russia and Alaska (Supplementary Fig. 2), and a separate small clade containing *A. kamchatica* subsp. *kawasakiana* accessions along with a few divergent accessions of *A. kamchatica* subsp. *kamchatica*. However, the relationship between these clades is different between the subgenomes. The clade containing subsp. *kawasakiana* is sister to the large Japanese in the *lyrata*-derived subgenome and it is sister to the Russia/Alaska clade in the *halleri*-derived subgenome (Supplementary Fig. 3). Different structure assignments and phylogenetic branching patterns between the subgenomes is not compatible with the scenario of a single origin of polyploidization, and supports that multiple parental individuals contributed to the origin of *A. kamchatica*.

Homeolog specific PCR

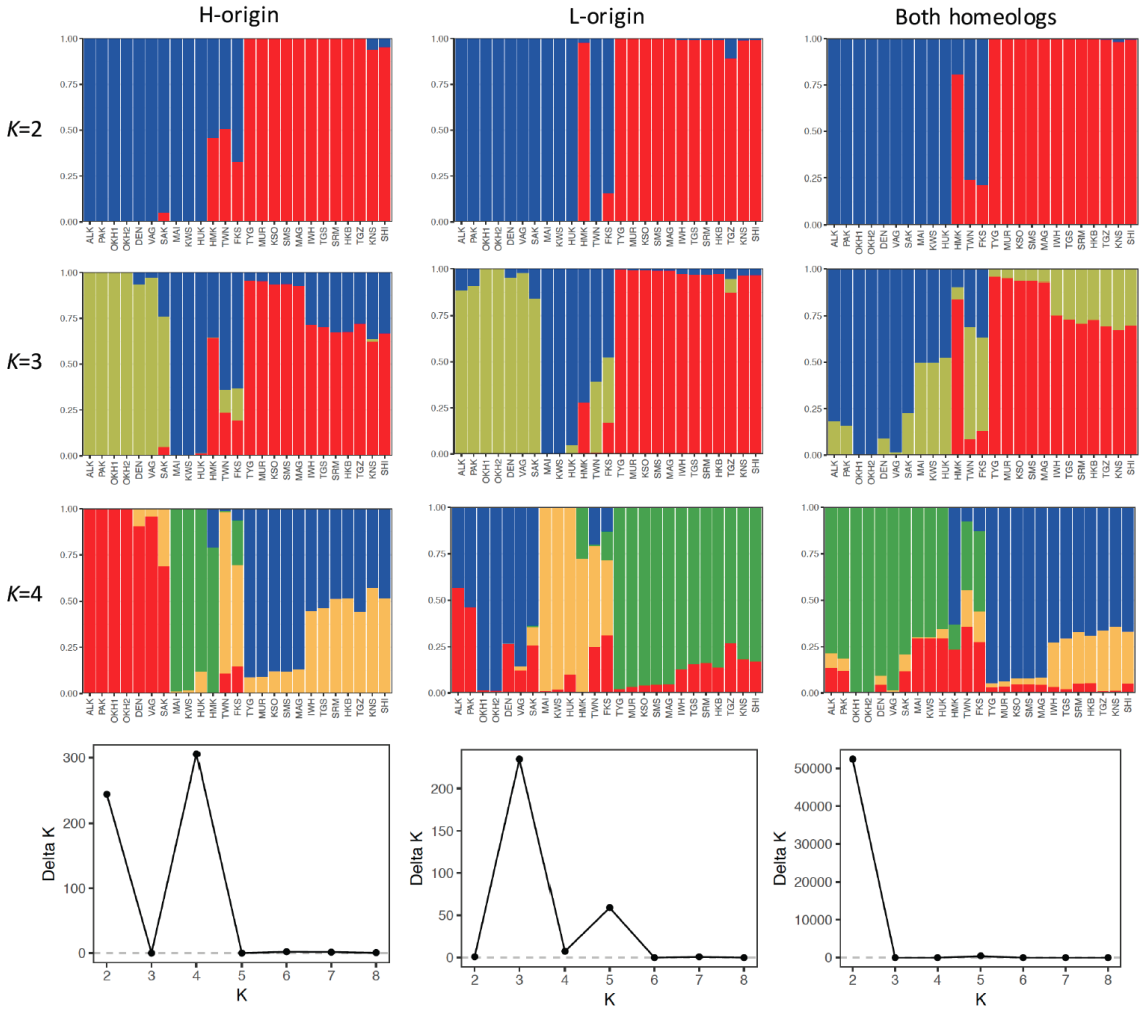
We performed Sanger sequencing using homeolog specific PCR to validate the read sorting method using *halleri* or *lyrata*-origin SNPs for the following genes (TAIR10 IDs): AT1G02180, AT1G02290, AT1G02630 (lyrata only), AT1G17770, AT3G17360, AT3G10570, AT3G17611, AT4G01860 (lyrata

only), AT4G26610, AT4G36080 (KWS halleri only), AT5G13930: CHS, AT5G14750: WER. Sequence fragments ranged from 170 bp to 1500 bp comprising a total of ca. 10 kb in length for the MUR, PAK and KWS accessions (OHK accession was used for WER *halleri*-homeolog). We defined SNP positions based on differences between homeologous regions, where sequences were often enriched for SNPs due to highly divergent intron polymorphisms. The alignments consisted of 285 divergent positions between the two homeologs. Because most of these positions were represented multiple (up to three) *A. kamchatica* individuals, we counted a total of 1375 sites. Among these only three SNPs were present in Sanger sequences that were different than the NGS data out of 1375 total SNPs. However, the other SNPs in these sequences corresponded perfectly to their respective homeologous sequences and therefore still validated the read sorting method. We also had cases where double peaks were present in the Sanger sequences for one of the two homeologs, but in all cases the two SNPs correspond to those shown in the NGS data for both homeologs, so both homeologs were partially amplified. We nevertheless consider these cases as supporting the NGS data since one homeolog was supported by Sanger data and both alleles were present in the other sequences.

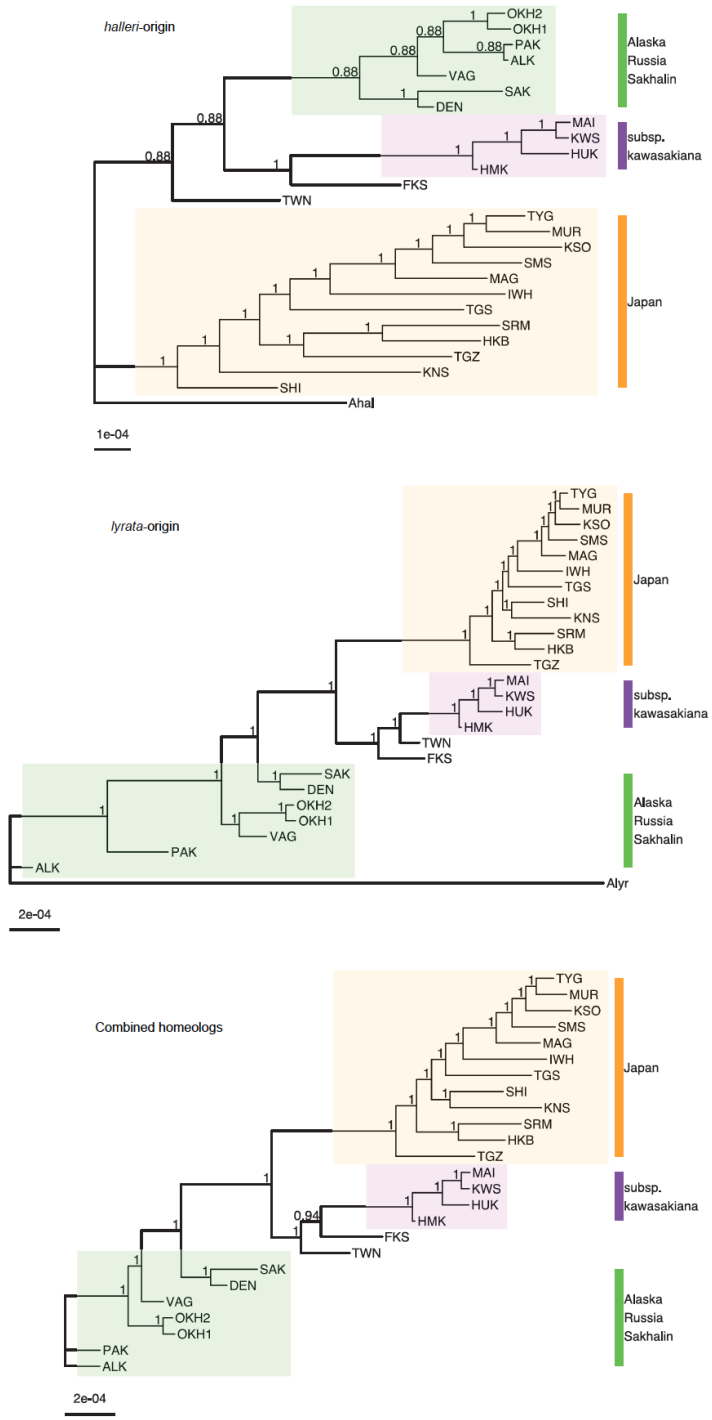
Supplementary Figures



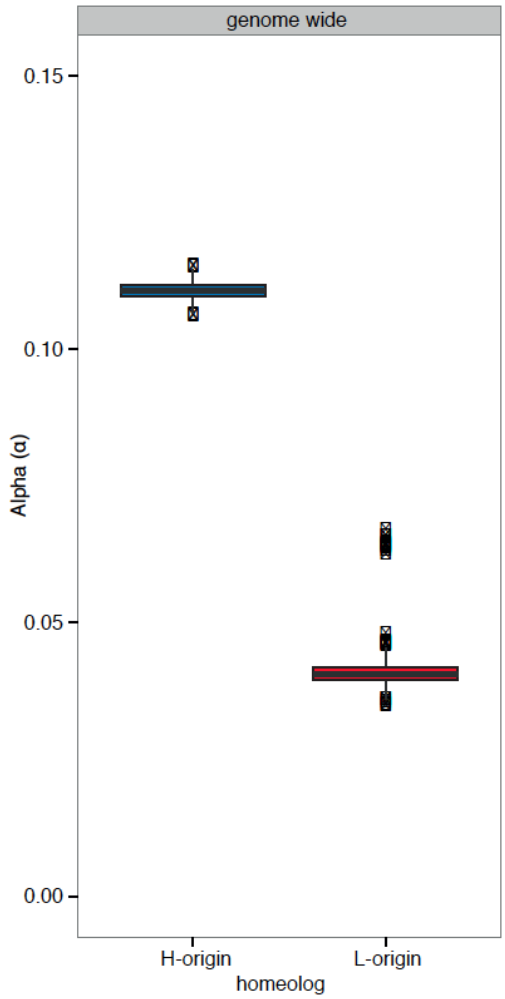
Supplementary Fig. 1. Linkage disequilbrium of *halleri*-origin (A) and *lyrata*-origin (B) subgenomes using 1 Mb windows along scaffolds. The blue (A) and red (B) curves represent the mean LD decay, while gray region is 50% confidence interval, and blue region is 90% confidence interval surrounding the means. The mean *lyrata*-origin LD remains at 0.47 while *halleri*-origin LD levels off at 0.34 at the scale of 1Mb genomic regions.



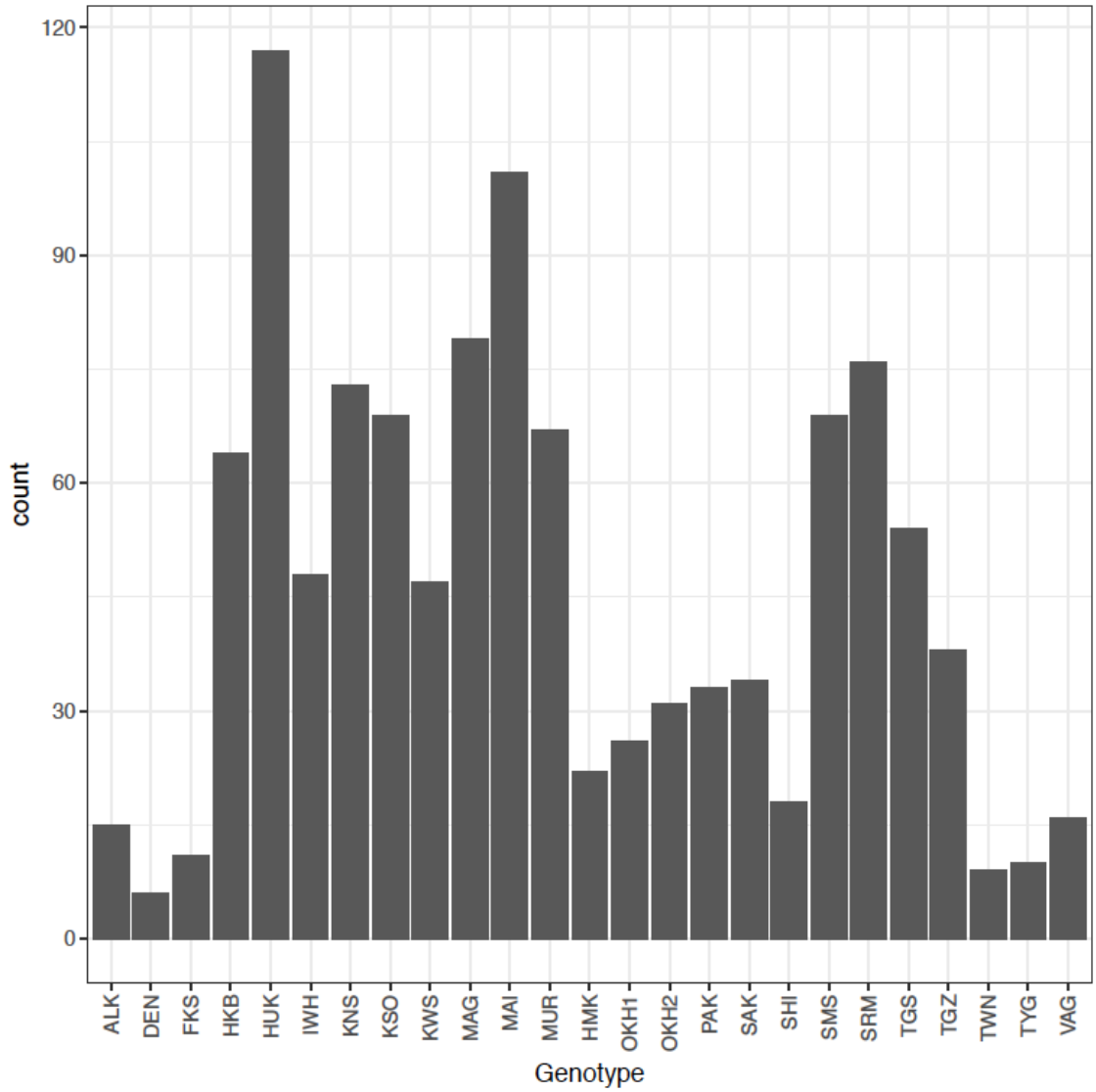
Supplementary Fig. 2. STRUCTURE assignments of *halleri* (H-origin) and *lyrata* (L-origin) derived homeologs for 25 *A. kamchatica* accessions for *K* = 2 to *K* = 4. The third column is the STRUCTURE assignments using SNPs from both homeologs combined. The Delta K [19] plots show the most likely *K* group clustering to be *K* = 4 for H-origin, *K* = 3 for L-origin and *K* = 2 using SNPs from both homeologs.



Supplementary Fig. 3. Phylogenetic relationships of 25 *A. kamchatica* accessions (top: *halleri*-genome; middle: *lyrata*-genome; bottom: both homeologs combined). Homeologs specific trees show clustering of a large clade of Japanese accessions (orange), and a distinct clade of northern latitude accessions (green) that are all *A. kamchatica* subsp. *kamchatica*. The small clustering of the *A. kamchatica* subsp. *kawasakiana* accessions is shown in purple, and is sister to the Japan clade in the *lyrata*-derived phylogeny, but sister to the Alaska/Russia in *halleri*-derived phylogeny. One accession from Taiwan is basal to the *kawasakiana* clade, and this lineage also contains an accession from Fukushima, Japan (FKS).



Supplementary Fig. 4. Estimates of adaptive evolution with all 25 *A. kamchatica* accessions. Mean α for H-origin was 0.11 (CI:1.08e-1 , 1.14e-1) and for L-origin α was 0.04 (CI: 3.68e-2 ,4.45e-2). CI are 95% confidence intervals.



Supplementary Fig. 5. Frequencies of genes with high-impact mutations in each genotype when both homeologs have disruptive mutations (the distribution of 511 genes from Supplementary Table 7 below).

Supplementary Tables

Supplementary Table 1. Gene annotation of v2.2 of Siberian *A. lyrata* subsp. *petraea* and v2.2 of *A. halleri* subsp. *gemmifera* (Tada Mine), Joint Genome Inititiave (JGI) *A. lyrata* subsp. *lyrata*, and *A. thaliana.* Annotation of *A. halleri* and *A. lyrata* v2.2 assembly produced with RNAseq hints.

Annotation	Genes	mRNA	Exons
A. halleri v2.2	32,553	34,553	187,838
A. lyrata v2.2	31,232	33,157	181,219
A. lyrata JGI[a]	32,670	32,670	NA
A. thaliana[b]	28,775	35,386	215,909

a: A. lyrata MN47 v1.07 assembly genome annotation from[13] MN47 v1.07 genome assembly by Hu et al.[16]

b: *A. thaliana* genome annotation from TAIR10.

Supplementary Table 2. Reciprocal best BLAST hits among three *Arabidopsis* species genome assemblies using our v.2.2 gene annotations in Supplementary Table 1. Only the longest transcript per gene was selected for the analysis. Hits A on B: hits from BLAST alignment of genes from the gene annotation A against the gene annotation B; RBH: reciprocal best BLAST hits. *A. lyrata* MN47 v1.07 assembly genome annotation from Hu et al. (2011) available from JGI. *A. thaliana* TAIR10 annotation.

Annotation A	Annotation B	Hits A on B	Hits B on A	RBH
A. halleri v2.2	*A. lyrata* v2.2	28,728	27,895	23,529
A. halleri v2.2	*A. thaliana*	25,328	23,728	21,433
A. halleri v2.2	*A. lyrata* JGI	26,402	26,917	22,447
A. lyrata v2.2	*A. lyrata* JGI	25,820	26,985	22,894
A. lyrata v2.2	*A. thaliana*	24,689	23,720	21,472
A. thaliana	*A. lyrata* JGI	24,033	25,716	21,941

Supplementary Table 3. List of 25 *A. kamchatica* accessions, sampling locations and sequencing depth. (See accompanying Excel file).

Accession	Species	location	lat_lon	Reads	otal_Coverage	Sorted_Ahal	Sorted_Alyr	orted_Common	Total_Ahal	Total_Alyr	Cov_Ahal	Cov_Alyr
ALK	Arabidopsis kamchatica subsp. kamchatica	USA: Alaska, Richardson Highway, South of Darling Creek bridge	63.4N 145.8W	41,222,868	8.8	14,118,890	10,796,478	1,121,305	15,172,336	11,851,960	7.8	6.8
DEN	Arabidopsis kamchatica subsp. kamchatica	Russia: Kamchatskii krai, near the river Denokhonok	54.2N 158.1E	29,858,928	6.3	10,393,815	7,703,567	764,879	11,122,492	8,427,447	5.7	4.9
FKS	Arabidopsis kamchatica subsp. kamchatica	Japan: Fukushima	37.2N 139.9E	40,709,648	8.7	14,460,719	10,002,155	1,729,908	16,138,372	11,676,183	8.3	6.7
HKB	Arabidopsis kamchatica subsp. kamchatica	Japan: Hakubayari	36.7N 137.8E	83,323,614	17.7	29,795,327	22,072,831	2,089,056	31,789,560	24,049,797	16.4	13.9
HMK	Arabidopsis kamchatica subsp. kawasakiana	Japan: Toyama	36.7N 137.3E	70,437,142	15	23,995,643	17,543,808	1,878,282	25,794,172	19,323,093	13.3	11.1
HUK	Arabidopsis kamchatica subsp. kawasakiana	Japan: Hukiinoura	34.6N 136.6E	106,898,796	22.7	37,136,448	27,677,971	2,750,024	39,756,647	30,274,808	20.5	17.5
IWH	Arabidopsis kamchatica subsp. kamchatica	Japan: Iwahana	35.9N 137.8E	52,549,196	11.2	18,872,866	14,148,184	1,425,613	20,239,313	15,500,712	10.4	8.9
KNS	Arabidopsis kamchatica subsp. kamchatica	Japan: Kinasa	36.7N 138.0E	91,705,866	19.5	32,746,897	24,388,091	2,336,948	34,975,970	26,595,190	18	15.3
KSO	Arabidopsis kamchatica subsp. kamchatica	Japan: Kisokomagatake	35.8N 137.8E	82,333,488	17.5	29,916,953	22,331,282	2,216,628	32,037,117	24,432,965	16.5	14.1
KWS	Arabidopsis kamchatica subsp. kawasakiana	Japan: Takashima	35.4N 136.0E	57,556,260	12.2	18,879,180	13,938,194	1,710,830	20,528,832	15,573,108	10.6	9
MAG	Arabidopsis kamchatica subsp. kamchatica	Japan: Magosazima	35.3N 138.3E	100,480,490	21.4	35,909,855	26,847,065	2,604,139	38,400,425	29,310,981	19.8	16.9
MAI	Arabidopsis kamchatica subsp. kawasakiana	Japan: Maiamihama	35.1N 136.0E	99,558,822	21.2	34,004,362	25,237,282	2,475,873	36,363,640	27,575,460	18.7	15.9
MUR	Arabidopsis kamchatica subsp. kamchatica	Japan: Murodo	36.2N 137.4E	92,156,804	19.6	28,450,072	20,969,486	1,884,357	30,239,914	22,723,204	15.6	13.1
OKH1	Arabidopsis kamchatica subsp. kamchatica	Russia: Khabarovsk krai, Okhotskii raion, foothills of Mt. Lanzhinskie gory	59.4N 143.3E	47,972,636	10.2	16,240,333	12,120,236	1,057,909	17,225,075	13,098,741	8.9	7.6
OKH2	Arabidopsis kamchatica subsp. kamchatica	Russia: Khabarovsk krai, Okhotskii raion, foothills of Mt. Lanzhinskie gory	59.4N 143.3E	45,097,050	9.6	15,717,431	11,766,389	1,179,854	16,842,632	12,882,931	8.7	7.4
PAK	Arabidopsis kamchatica subsp. kamchatica	USA: Alaska, Potter	61.2N 149.3W	180,109,152	38.3	27,117,743	20,836,034	2,642,737	29,631,690	23,318,884	15.3	13.4
SAK	Arabidopsis kamchatica subsp. kamchatica	Russia: Sakhalin, Marakovskii raion, Zaozernoye	48.4N 142.7E	51,986,358	11.1	16,978,097	12,521,408	1,214,392	18,133,252	13,663,380	9.3	7.9
SHI	Arabidopsis kamchatica subsp. kamchatica	Japan: Shirakawa	36.3N 136.9E	36,818,198	7.8	12,491,175	9,099,814	863,158	13,310,606	9,909,963	6.9	5.7
SMS	Arabidopsis kamchatica subsp. kamchatica	Japan: Simasimadani	36.2N 137.8E	82,024,638	17.4	28,623,779	21,103,492	2,121,047	30,651,632	23,110,874	15.8	13.3
SRM	Arabidopsis kamchatica subsp. kamchatica	Japan: Mt. Shirouma	36.8N 137.8E	103,381,614	22	37,373,380	27,889,832	2,881,603	40,139,114	30,636,345	20.7	17.7
TGS	Arabidopsis kamchatica subsp. kamchatica	Japan: Mt. Shikokutsurugi	33.9N 134.1E	64,747,706	13.8	23,060,935	16,670,868	1,747,686	24,729,621	18,328,385	12.7	10.6
TGZ	Arabidopsis kamchatica subsp. kamchatica	Japan: Tsurugigozen	36.6N 137.6E	49,130,130	10.4	17,080,259	12,600,990	1,278,910	18,304,390	13,814,192	9.4	8
TWN	Arabidopsis kamchatica subsp. kamchatica	Taiwan: Taroko national park	24.0N 121.3E	27,214,840	5.8	9,374,591	6,723,726	725,941	10,068,892	7,412,536	5.2	4.3
TYG	Arabidopsis kamchatica subsp. kamchatica	Japan: Tateyamagawa	36.6N 137.6E	33,247,108	7.1	11,084,924	8,145,924	755,793	11,802,582	8,856,608	6.1	5.1
VAG	Arabidopsis kamchatica subsp. kamchatica	Russia: Kamchatskii krai, near the river Vaktan Ganal'skii	53.5N 157.6E	40,538,374	8.6	13,933,714	10,460,424	960,743	14,843,746	11,362,106	7.6	6.6

Supplementary Table 4. Diversity and polymorphism statistics of both subgenomes by sliding window analysis.

Ahal whole genome	bases	Bases (%)	polym	Polym	nucDiv	θ_w	π
overall	163517656	1	1138032	0.007	0.0017	0.0018	0.0017
gene	75291060	0.4604	454338	0.006	0.0015	0.0016	0.0015
coding	38896876	0.2379	216194	0.0056	0.0014	0.0015	0.0014
intron	22946734	0.1403	154633	0.0067	0.0017	0.0017	0.0017
intergenic	83592223	0.5112	660511	0.0079	0	0.0035	0.0033

Alyr whole genome	bases	Bases (%)	polym	Polym	nucDiv	θ_w	π
overall	149864674	1	946600	0.0063	0.0017	0.0017	0.0017
gene	72299008	0.4824	436107	0.006	0.0016	0.0016	0.0016
coding	37093072	0.2475	205023	0.0055	0.0015	0.0015	0.0015
intron	21685851	0.1447	146380	0.0068	0.0018	0.0018	0.0018
intergenic	74042836	0.4941	496233	0.0067	0	0.0034	0.0034

Supplementary Table 5. Estimated effective (N_e) population sizes using empirical diversity estimates and published mutation accumulation rates. The calculation of $N_e = \pi_{syn}/4\mu$ using the mutation rate from Koch et al.[28] used only synonymous diversity while the calculation using Ossowski et al.[29] used total diversity.

Species	π_{syn}	π_{total}	N_e (1)	N_e (2)
A. kamchatica	0.0046	0.0015	77000	53571
A. halleri	0.028	0.0097	466667	364202
A. lyrata	0.029	0.0102	483333	345041
(1) Koch et al.[28] mutation rate μ	1.50E-08			
(2) Ossowski et al.[29] mutation rate μ	7.00E-09			

Supplementary Table 6. Samples used for estimating DFE and α in Fig. 4A and E in main text. Illumina reads from European *A. halleri* and *A. lyrata* were obtained from[30]. SNPs in diploid parents were phased and separated into two alleles, indicated by _1 and _2 following accession number. To get equal sample size, *A. lyrata* alleles and *A. kamchatica* samples were chosen at random.

A. kamchatica	***A. lyrata***	***A. halleri***
ALK	SRR2040790_1	SRR2040780_1
DEN	SRR2040791_2	SRR2040780_2
HKB	SRR2040792_1	SRR2040782_1
IWH	SRR2040793_2	SRR2040782_2
KNS	SRR2040794_1	SRR2040783_1
KSO	SRR2040795_1	SRR2040783_2
MAG	SRR2040795_2	SRR2040784_1
MUR	SRR2040796_2	SRR2040784_2
OKH1	SRR2040797_2	SRR2040785_1
OKH2	SRR2040798_1	SRR2040785_2
PAK	SRR3111438_2	SRR2040786_1
SAK	SRR3111439_1	SRR2040786_2
SHI	SRR3111439_2	SRR2040787_1
SMS	SRR3111440_1	SRR2040787_2
SRM	SRR3111441_1	SRR2040810_1
TGS	SRR3111441_2	SRR2040810_2
TGZ	SRR3111442_2	SRR3107262_1
VAG	SRR3111443_1	SRR3107262_2

Supplementary Table 7. High impact mutations. Counts are the number of homeologs with one or more of any of the mutation types.

Homeolog	frameshift variant	start lost	stop gained	stop lost	total[a]	% total
H-origin	3311	282	1662	190	4219	20.78
L-origin	4014	423	2002	251	4952	24.39
Shared in both homeologs[b]					1559	7.68
Shared in genotypes[c]					511	2.52

a: total number of homeologs with one or more high impact mutations (multiple mutation types are possible in a single homeolog)

b: total number of genes with high impact mutations in both homeologs out of 25 individuals

c: total number of high impact mutations in both homeologs in a single individual

Supplementary Table 8. Gene Ontology of high impact mutations for premature stop codon and frameshift combined for *halleri* (A) and *lyrata* (B) derived coding sequences. (C) GO analysis of genes with any of the four high impact mutation types (from Supplementary Table 6) where both homeologs in a single genotype had disruptive mutations.

H-origin								
GO_acc	term_type	Term	queryitem	querytotal	refitem	reftotal	pvalue	FDR
GO:0003824	F	catalytic activity	1507	4273	6350	19936	6.90E-05	0.036
GO:0016787	F	hydrolase activity	567	4273	2285	19936	0.00014	0.036
GO:0001883	F	purine nucleoside binding	260	4273	983	19936	0.00015	0.036
GO:0001882	F	nucleoside binding	260	4273	983	19936	0.00015	0.036
GO:0030554	F	adenyl nucleotide binding	260	4273	983	19936	0.00015	0.036
GO:0019825	F	oxygen binding	58	4273	159	19936	1.10E-05	0.011
GO:0012501	P	programmed cell death	45	4273	111	19936	4.60E-06	0.012
GO:0008236	F	serine-type peptidase activity	42	4273	115	19936	0.00016	0.036
GO:0017171	F	serine hydrolase activity	42	4273	115	19936	0.00016	0.036
GO:0006915	P	apoptosis	32	4273	61	19936	1.10E-07	0.00056
GO:0004888	F	transmembrane receptor activity	31	4273	64	19936	1.60E-06	0.0033
L-origin								
GO_acc	term_type	Term	queryitem	querytotal	refitem	reftotal	pvalue	FDR
GO:0016787	F	hydrolase activity	663	5031	2285	19936	5.90E-05	0.022
GO:0017076	F	purine nucleotide binding	346	5031	1146	19936	0.00013	0.03
GO:0001882	F	nucleoside binding	314	5031	983	19936	2.50E-06	0.0019
GO:0001883	F	purine nucleoside binding	314	5031	983	19936	2.50E-06	0.0019
GO:0030554	F	adenyl nucleotide binding	314	5031	983	19936	2.50E-06	0.0019
GO:0032559	F	adenyl ribonucleotide binding	294	5031	927	19936	8.80E-06	0.0044
GO:0005524	F	ATP binding	292	5031	921	19936	9.70E-06	0.0044
GO:0017111	F	nucleoside-triphosphatase activity	185	5031	574	19936	0.00013	0.03
GO:0019825	F	oxygen binding	61	5031	159	19936	0.00019	0.04
GO:0008236	F	serine-type peptidase activity	48	5031	115	19936	8.40E-05	0.024
GO:0017171	F	serine hydrolase activity	48	5031	115	19936	8.40E-05	0.024
Shared in both homeologs in a single genotype								
GO_acc	term_type	Term	queryitem	querytotal	refitem	reftotal	pvalue	FDR
GO:0060089	F	molecular transducer activity	17	497	239	19936	0.00014	0.013
GO:0004871	F	signal transducer activity	17	497	239	19936	0.00014	0.013
GO:0004872	F	receptor activity	11	497	95	19936	2.90E-05	0.0052
GO:0012501	P	programmed cell death	11	497	111	19936	0.00012	0.042
GO:0004888	F	transmembrane receptor activity	10	497	64	19936	4.50E-06	0.0016
GO:0006915	P	apoptosis	9	497	61	19936	2.10E-05	0.015

Supplementary References

1. Shimizu-Inatsugi, R. *et al.* The allopolyploid *Arabidopsis kamchatica* originated from multiple individuals of *Arabidopsis lyrata* and *Arabidopsis halleri*. *Mol. Ecol.* **18,** 4024–4048 (2009).
2. Schmickl, R., Jørgensen, M. H., Brysting, A. K. & Koch, M. A. The evolutionary history of the Arabidopsis lyrata complex: a hybrid in the amphi-Beringian area closes a large distribution gap and builds up a genetic barrier. *BMC Evol. Biol.* **10,** 98 (2010).
3. Tsuchimatsu, T. *et al.* Evolution of self-compatibility in Arabidopsis by a mutation in the male specificity gene. *Nature* **464,** 1342–1346 (2010).
4. Akama, S., Shimizu-Inatsugi, R., Shimizu, K. K. & Sese, J. Genome-wide quantification of homeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid Arabidopsis. *Nucleic Acids Res.* **42,** e46–e46 (2014).
5. Briskine, R. V. *et al.* Genome assembly and annotation of *Arabidopsis halleri* , a model for heavy metal hyperaccumulation and evolutionary ecology. *Mol. Ecol. Resour.* (2016). doi:10.1111/1755-0998.12604
6. Butler, J. *et al.* ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Res.* **18,** 810–820 (2008).
7. AUGUSTUS Development Team. Incorporating RNAseq data into AUGUSTUS with TopHat. (2014). Available at: http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=IncorporatingRNAseq.Tophat. (Accessed: 15th January 2015)
8. Paape, T. *et al.* Conserved but Attenuated Parental Gene Expression in Allopolyploids: Constitutive Zinc Hyperaccumulation in the Allotetraploid *Arabidopsis kamchatica*. *Mol. Biol. Evol.* **33,** 2781–2800 (2016).
9. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
10. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-3.0. (1996). Available at: http://www.repeatmasker.org.
11. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
12. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485,** 635–641 (2012).

13. Rawat, V. *et al.* Improving the Annotation of Arabidopsis lyrata Using RNA-Seq Data. *PLOS ONE* **10,** e0137391 (2015).
14. Roux, C. *et al.* Does Speciation between Arabidopsis halleri and Arabidopsis lyrata Coincide with Major Changes in a Molecular Target of Adaptation? *PLoS ONE* **6,** e26872 (2011).
15. Al-Shehbaz, I. A. & O'Kane, S. L. Taxonomy and Phylogeny of Arabidopsis (Brassicaceae). *Arab. Book* **1,** e0001 (2002).
16. Hu, T. T. *et al.* The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43,** 476–481 (2011).
17. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12,** 656–664 (2002).
18. Hanikenne, M. *et al.* Evolution of metal hyperaccumulation required cis-regulatory changes and triplication of HMA4. *Nature* **453,** 391–395 (2008).
19. Arrivault, S., Senger, T. & Krämer, U. The Arabidopsis metal tolerance protein AtMTP3 maintains metal homeostasis by mediating Zn exclusion from the shoot under Fe deficiency and Zn oversupply. *Plant J.* **46,** 861–879 (2006).
20. Filatov, V. *et al.* Comparison of gene expression in segregating families identifies genes and genomic regions involved in a novel adaptation, zinc hyperaccumulation: GENE EXPRESSION IN SEGREGATING FAMILIES. *Mol. Ecol.* **15,** 3045–3059 (2006).
21. Talke, I. N. Zinc-Dependent Global Transcriptional Control, Transcriptional Deregulation, and Higher Gene Copy Number for Genes in Metal Homeostasis of the Hyperaccumulator Arabidopsis halleri. *PLANT Physiol.* **142,** 148–167 (2006).
22. Shahzad, Z. *et al.* The Five AhMTP1 Zinc Transporters Undergo Different Evolutionary Fates towards Adaptive Evolution to Zinc Tolerance in Arabidopsis halleri. *PLoS Genet.* **6,** e1000911 (2010).
23. Lamesch, P. *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40,** D1202–D1210 (2012).
24. Shahzad, Z. *et al. Plant Defensin type 1* (*PDF1*): protein promiscuity and expression variation within the *Arabidopsis* genus shed light on zinc tolerance acquisition in *Arabidopsis halleri*. *New Phytol.* **200,** 820–833 (2013).
25. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).

26. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol. Ecol.* **14,** 2611–2620 (2005).

27. Tsuchimatsu, T., Kaiser, P., Yew, C.-L., Bachelier, J. B. & Shimizu, K. K. Recent Loss of Self-Incompatibility by Degradation of the Male Component in Allotetraploid Arabidopsis kamchatica. *PLoS Genet.* **8,** e1002838 (2012).

28. Koch, M. A., Haubold, B. & Mitchell-Olds, T. Comparative Evolutionary Analysis of Chalcone Synthase and Alcohol Dehydrogenase Loci in Arabidopsis, Arabis, and Related Genera (Brassicaceae). *Mol. Biol. Evol.* **17,** 1483–1498 (2000).

29. Ossowski, S. *et al.* The Rate and Molecular Spectrum of Spontaneous Mutations in Arabidopsis thaliana. *Science* **327,** 92–94 (2010).

30. Novikova, P. Y. *et al.* Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.* **48,** 1077–1082 (2016)